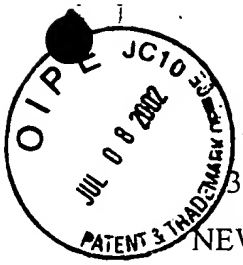

BAKER & BOTTS, L.L.P.
30 ROCKEFELLER PLAZA
NEW YORK, NEW YORK 10112

TO ALL WHOM IT MAY CONCERN:

Be it known that WE, ANDREY RZHETSKY and SERGEY KALACHIKOV, citizens of Russia, whose post office addresses are 560 Riverside Drive, 11F New York, New York 10027; and 154 Haven Avenue, 1303, New York, New York 10032 respectively; MICHAEL O. KRAUTHAMMER, citizen of Switzerland, whose post office address is 27 W. 76th Street, Apt. 3A, New York, N.Y., 10023; CAROL FRIEDMAN and PAULINE KRA, citizens of the United States, whose post office addresses are 14 Dimitri Place, Larchmont, New York, 10538 and 109-14 Ascan Ave. Forest Hills, N.Y., 11375, respectively, have invented an improvement in

Methods for Extracting Information on Interactions Between Biological Entities From Natural Language Text Data.

of which the following is a

SPECIFICATION

This application is a continuation-in-part of pending application Serial No. 09/327,938 filed June 8, 1999 which claims priority to provisional patent application Serial No. 60/129,469 filed April 15, 1999. The invention described herein was funded in part by a grant from the National Library of Medicine, namely, Grant Number's LM06274 and LM05627. The United States Government may have certain rights to the invention. The present specification contains a computer program listing which appears as a microfiche Appendix H.

5.3. GENERATION OF SPECIALIZED DATABASES

In accordance with the present invention, specialized databases may be developed that contain information derived from unpublished data, publications such as research articles, theses, posters, abstracts, etc. and/or databases concerning interactions among genes and proteins, their domain/motif structure, and their biological functions.

For example, but not by way of limitation, a specialized database may be prepared as follows. Protein and gene sequences may be provided, for example, by the Java program PsiRetrieve which allows for quick retrieval of protein or nucleotide sequences from binary BLAST databases by sequence accession number, keyword or groups of keywords, or species name. In addition, using the program PsiRetriever, sequences encoding the proteins of interest may be retrieved from the non-redundant (NCBI) database of protein sequences and stored as a FASTA file. The FASTA file is then converted into a binary blast database using the program FORMATDB from the BLAST suit of programs.

Known motifs/domains for proteins may also be collected using the flat file versions of major protein databases, such as SwissProt and the non-redundant database of NCBI. The databases can be downloaded and searched for the keywords "motif" and "domain" in the feature tables of proteins. In addition, existing databases of motifs and

domains, such as BLOCKS and pfam, can be downloaded (Henikoff et al., 1991, NAR 19:6565-6572). Still further, it is understood that any publically available database containing gene/protein sequences may be utilized to generate the specialized databases for use in the practice of the present invention.

Homologous sequences may be aligned using, for example, the CLUSTALW program (Higgins, et al. 1996 Methods in Enzymology 266: 383-402). A protein's sequence corresponding to each domain/motif can be identified, saved and used for building a Hidden Markov Model (HMM) of the domain/motif using a HMMER and HMMER2 packages (see, Durbin, R. et al. 1998 in Biological Sequence Analysis: Probablistic Models of Proteins and Nucleic Acids). HMMER and HMMER2 packages are useful for (i) building HMMs from sets of aligned protein or nucleotide sequences, and (ii) comparing the HMMs with sequence databases aimed at identifying significant similarities of HMMs with database sequences. Both nucleotide and protein databases can be used for this purpose. Alternatives to the Hidden Markov Model method for building domain/motif models include neural network motif analysis (Wu, C.H. et al., 1996, Comput Appl Biosci 12, 109-18; Hirst, J.D., 1991, Protein Eng 4:615-23) and positional weight matrix analysis (Claverie, J.M., 1994, Comput Chem 18:287-94;

motivated by a probability model (Fitch, 1974 J. Mol. Evol. 3:263-268)), (ii) the maximum likelihood method (Goldman, 1990, Syst. Zool. 30:345-361; Yang et al., 1995, Syst. Biol. 44:384-399; Felsenstein, J., 1996, Methods Enzymol. 266-418-427); and (iii) the distance matrix tree making method (Saito, N. and Nei, M., 1987, Mol. Biol. Evol. 4:406-425). Since the data analyses of orthologs and paralogs often involve very distantly related sequences, the maximum likelihood method is preferably used for small data sets and the distance-matrix method in other instances.

To construct a reconciled tree according to the invention, the first step comprises a search for homologs in a publicly or privately available database such as, for example, GenBank, Incyte, binary BLAST databases, Swiss Prot and NCBI databases. Following the identification of homologous sequences a global alignment is performed using, for example, the CLUSTALW program. From the sequence alignment a gene tree is constructed using, for example, the computer program CLUSTLAW which utilizes the neighbor-joining method of Saito and Nei (1997, Mol. Biol. Evol. 4:406-425). Construction of a species tree is then retrieved from, for example, the following database 3.NCBI.NLM.NIH.GOV//taxomy.tax.html.

The species tree and gene tree are given as input into the algorithm described below, which integrates both trees into a reconciled tree. Agreement between the gene tree and the corresponding species tree for any given set of sequences indicates

6.1 APOPTOSIS GENE DISCOVERY METHOD

Identification of a putative apoptosis-related human gene began with an identification of all genes in C. elegans that contained either a POZ or kelch domain. A subset of these genes is shown in Figure 13. Hidden Markov Models (HMM) for the POZ and Kelch domains were built as follows. Starting with POZ and kelch sequences from the Drosophilia kelch protein (gi 577275) homologs were identified in other protein sequences using the BLASTP program. The resulting sequences showing significant similarity (e-value less than 0.001) were aligned using CLUSTALW program and the alignments were used to build Hidden Markov Models with HMMER-2 package (Krogh et al., 1995). A computer printout listing of HMM models of tumor suppressors appears as a Microfiche H to the present specification.

The resulting models were used to search through a database collection of C.elegans protein sequences. The domain structures of proteins having either a POZ or kelch domain were identified using existing collections of protein.